COMPARISON OF METHODS FOR DETECTING TREATMENT EFFECTS UNDER SKEWED DISTRIBUTIONS

SHARIPAH SOAAD SYED YAHAYA

UNIVERSITI UTARA MALAYSIA
2009

DISCLAIMER

I am responsible for the accuracy of all opinions, technical comments, factual reports, data, figure and illustrations in the article. I bear full responsibility for the checking whether material submitted is subjected to copyright or ownership right. Universiti Utara Malaysia (UUM) has no liability for the accuracy of such comments, reports, technical and factual information and the copyright or ownership right claims.

CHIEF RESEARCHER:

_____

SHARIPAH SOAAD SYED YAHAYA

DATE: 1 JANUARY 2009

ACKNOWLEDGEMENT

First and foremost, I would like to extend my sincere thanks to UUM, which funded this research. Appreciation also goes to my fellow colleagues and friends who had provided guidance and ideas as well as encouragement that has definitely encourage me to complete this research. Last but not least, special thanks to all those who had lend a helping hand in allowing me to materialize this research.

Sharipah Soaad Syed Yahaya

TABLE OF CONTENTS

1.0 BACKGROUND OF STUDY

The two general problems that researchers always encounter when using the traditional methods in detecting treatment effects are non-normality and heteroscedasticity. For example, when using analysis of variance (ANOVA), these problems can seriously hamper the ability to detect true differences between means by causing the Type I error rates to inflate from the nominal value. This will result in spurious rejections of the null hypotheses of equal means. Nonparametric counterparts of those procedures, namely the Kruskal-Wallis was developed to deal with such problems. However, this non parametric procedures is more appropriate for nonnormal but symmetric data. Furthermore nonparametric procedures are less powerful than parametric procedures; therefore, require a larger sample size to reject a false hypothesis. In practice, we often need to estimate location and/or scale from small sample. The sample size n is often constrained by the cost of an observation. In many experimental settings (e.g. in chemistry, biology, medical) one will typically repeat each measurement only a few times. Even a small sample may contain aberrant values due to technical problems or measurement inaccuracies. Since the sample is small, getting rid off the aberrant values should be avoided. If this is the case, a researcher would need an estimator that is stable and insensitive to these aberrant values or inaccuracies. This means that the estimator has to be robust.

In recent years, numerous methods for locating treatment effects by controlling Type I error to detect treatment effects are being studied. Progress has been made in terms of finding better methods for controlling Type I error to detect treatment effects in the one way independent group designs (Babu et al., 1999;

Othman et al., 2004; Wilcox & Keselman, 2003). Through a combination of impressive theoretical developments, more flexible statistical methods, and faster computers, serious practical problems that seemed insurmountable only a few years ago can now be addressed. These developments are important to applied researchers because they greatly enhance the ability to discover true differences between treatment groups.

The parametric approach in detecting treatment effects continued to play a prominent role because of its capacity to comprehensively describe information contained in a data. However, the good performance and valid application of the procedures require strict adherence to certain assumptions, which do not always operate as predicatively as assumed in the real world. Some of the most common statistical procedures are extremely sensitive to these minor deviations from assumptions such as in the case of normality of distributions and homogeneity of variances**.** As an example, when computing confidence intervals and testing hypothesis about means, the methods are based on the assumption that observations are randomly sampled from normal distributions. Another instance is when comparing dependent groups; where the methods are also assume that groups have a common variance. Currently, these methods form the backbone of most applied research that involves statistical methodology. It is therefore desirable to construct methods of inference that do not depend on distributional and homoscedasticity (equal variances) assumptions for their validity.

Consequently, non-parametric statistics emerged as a field of research and some of its methods become widely popular in applications. The basic principle was to make as few assumptions about the data as possible and still get the answer to a

specific question. However, non parametric procedures are more appropriate for data based on weak measurement scales.

In view of all the aforementioned violations, an estimator that is stable and insensitive to all these violations is needed. In other words, the estimator has to be robust. Robust statistics combine the virtues of both, the parametric and the non-parametric approach**.** In non-parametric inference, few assumptions are made regarding the distribution from which the observations are drawn. In contrast, the approach in robust inference is different wherein there is a working assumption about the form of the distribution, but we are not entirely convinced that the assumption is true. Robustness theories can be viewed as stability theories of statistical inference. What is desired is an inference procedure, which in some sense does almost as well as possible if the assumption is true, but does not perform much worse within a range of alternatives to the assumption. In order to achieve a good test, one needs to be able to control Type I error.

In their efforts to control the Type I error rate, investigators looked into numerous robust methods since these methods generally are insensitive to assumptions about the overall nature of the data. Any small deviations from the model assumptions should only slightly impair the performance, for example, the level of a test should be close to the nominal value calculated at the model, and larger deviations from the model should not cause catastrophe**.** Robust measures of central tendency such as trimmed means, medians or *M*-estimators have been considered as alternatives for the usual least squares estimator, i.e., the usual least squares means, in most research recently (e.g. Keselman et al, 2004b; Wilcox and Keselman, 2002; Luh and Guo, 1999; Wilcox et al. 1998). These measures of central tendency had been shown to have better control over Type I error (see e.g. Lix and Keselman,1998;

Othman et al., 2004; Wilcox, 1997; Yuen, 1974). Other investigators, e.g. Babu et al. (1999) used median as the central tendency measure when dealing with skewed distribution and Wilcox and Keselman (2003) introduced a modified one-step *M*-estimator (*MOM*) as the central tendency measure when testing for treatment effects. Among the latest procedures in robust statistics are the modified S_1 (Babu et al, 1999) and *MOM-H* (Othman et al, 2004).

1.1 S_1 Statistic

In the quest for a good robust statistics for testing location parameters for skewed distributions, Babu et al. (1999) discovered the S_1 statistics which uses the median as the central measures. It is the sum of all possible differences of sample medians from the *J* distributions divided by their respective sample standard errors.

Let $Y_{ij} = (Y_{1j}, Y_{2j}, ..., Y_{n_j j})$ be a sample from an unknown distribution F_j and let M_i be the population median of F_j : j = 1, 2, . . . , J. For testing $H_0 : M_1 = M_2 = \ldots = M_J$ versus $H_1 : M_i \neq M_j$ for at least one pair (i,j), the S_1 statistic is defined as

$$S_1 = \sum_{1 \le i < j \le J} |s_{ij}| \qquad [1.1]$$

where $s_{ij} = \dfrac{(\hat{M}_i - \hat{M}_j)}{\sqrt{(\hat{\omega}_i + \hat{\omega}_j)}}$,

$$\omega_j = \left(\frac{1}{n_i} \sum |Y_{ij} - \hat{M}_j| \right)^2 \qquad [1.2]$$

$$\hat{\omega}_j = \frac{\omega_j}{n_j}$$

$\hat{M}_j$ is the sample median from the jth group, of group j

ω_j is the squared mean absolute deviation from sample median $\hat{M}_j$, and

n_j is the sample size for group *j*.

1.2. *MOM-H* statistic

The *H* test is defined as

$$H = \frac{1}{N}\sum_{j=1}^{J} n_j(\hat{\theta}_j - \hat{\theta}.)^2, \text{ where} \qquad [1.3]$$

$$N = \sum_j n_j \text{ and}$$

$$\hat{\theta}. = \sum_j \hat{\theta}_j / J .$$

This statistic is readily adaptable to any measure of central tendency and appears to give reasonably good results when using the Harrel-Davis estimator of the median. However, its use is not recommended for the comparison of means or even trimmed means (Wilcox, 1997).

Othman et al. (2004) used the *H* test statistic when testing for the equality of the "typical" score across treatment groups. Nonetheless, they modified this statistic by replacing $\hat{\theta}$ with the *MOM* estimator (denoted as $\hat{\theta}_M$). The modified test statistic is known as *MOM-H*, and the goal of this statistic is to test $H_0: \theta_{M1} = \theta_{M2} = \ldots = \theta_{MJ}$ versus $H_1: \theta_{Mi} \neq \theta_{Mj}$ for at least one pair of (*i*, *j*). Othman and his colleagues found that *MOM-H* was quite effective in controlling rates of Type I error even though the data were heterscedastic and nonnormal in shape.

1.2.1 Modified one-step *M*- estimator (*MOM)*

MAD_n is the default scale estimator used in the criterion for determining extreme values in $\hat{\theta}_M$. (Refer to 1.3.1 for further discussion on MAD*n*).

Let $Y_j = (Y_{1j}, Y_{2j}, \ldots, Y_{nj})$ be a sample from an unknown skewed distribution F_j and let M_j be the population median of $F_j : j = 1, 2, \ldots, J$. The estimator as suggested by Wilcox and Keselman (2003) is defined as

$$\hat{\theta}_{M\,j} = \sum_{i=i_1+1}^{n_j - i_2} \frac{Y_{(i)\,j}}{n_j - i_1 - i_2} \qquad [1.4]$$

where

$Y_{(i)j}$ = the ith. ordered observations in group j.

n_j = # of observations for group j.

i_1 = # of observations Y_{ij} such that $(Y_{ij} - \hat{M}_j) < -2.24(MAD_{nj})$,

i_2 = # of observations Y_{ij} such that $(Y_{ij} - \hat{M}_j) > 2.24(MAD_{nj})$,

1.2.2 Criterion for Choosing the Sample Values

From Eqn. 1.4 the criterion used to determine the number of extreme observations in each group j, centers around the indices i_1 and i_2, where i_1 is the number of extreme observations in the left tail, while i_2 are the extreme observations in the right tail. For a sample with no extreme value, wherein $i_1 = i_2 = 0$, $\hat{\theta}_M$ is equal to the mean for the group. After eliminating the extreme values, calculate $\hat{\theta}_{M\,j}$ and proceed with the calculation of the H statistic.

Modification on S_1 and *MOM-H* is done by substituting the default scale estimator, $\hat{\omega}_j$ and MAD_n respectively with two of the most robust scale estimators known as S_n and T_n. This scale estimator was chosen based on its robustness properties such as highest breakdown point and bounded influence function. Breakdown point and influence function are the tools for judging robustness.

For the following sections, let $X = (x_1, x_2, \ldots, x_n)$ be a random sample from any distribution and let the sample median be denoted as $med_i x_i$.

1.3 Scale Estimators

The choice of scale estimators is important since these scale estimators will determine the performance of the central tendency measures especially in the *MOM-H* case where a measure of scale with a high breakdown point is needed in order for a central tendency measure is to have a high breakdown point as well.

1.3.1 *MADn*

MADn is median absolute deviation about the median. Given b

$$MADn = b\ med_i \left| x_i - med_j x_j \right| \qquad [1.5]$$

with *b* as a constant, this scale estimator is very robust with best possible breakdown point and bounded influence function. Huber (1981) identified *MADn* as the single most useful ancillary estimate of scale due to its high breakdown property. *MADn* is simple and easy to compute.

The constant *b* is needed to make the estimator consistent for the parameter of interest. For example if the observations are randomly sampled from a normal distribution, by including b = 1.4826, the *MADn* will estimate σ, the standard deviation. With constant b = 1, *MADn* will estimate 0.75σ, and this is known as *MAD*.

1.3.2 S_n

Rousseeuw and Croux (1993) suggested alternatives to *MADn* that can be used as intial or ancillary scale estimates that are more efficient and as well are not slanted towards symmetric distributions. One such estimator is S_n, defined as

$$S_n = c\ med_i \left\{ med_j \left| x_i - x_j \right| \right\} . \qquad [1.6]$$

This estimator is similar to *MADn*; the only difference between the two is that the med_j operation is transferred to the outside of the absolute value. This makes S_n a location free estimator. Another advantage is its explicit formula which means that this estimator is always uniquely defined. A modest simulation study by Rousseeuw and Croux (1993) found that a correction factor, c = 1.1926, succeeded in making S_n unbiased for finite samples. They also proved that S_n has the highest possible breakdown point. In terms of efficiency, S_n was proven to be more efficient (58.23%) than *MADn* (36.74%) with Gaussian distributions.

1.3.3 T_n

Another promising scale estimator proposed by Rousseeuw and Croux (1993) which possesses the attractive properties of the robust scale estimator is T_n defined as

$$T_n = 1.3800 \frac{1}{h} \sum_{k=1}^{h} \left\{ med_{j \neq i} \left| x_i - x_j \right| \right\}_{(k)} \qquad [1.7]$$

It was proven that T_n has the 50% breakdown point, a continuous influence function, and an efficiency of 52%, which makes it more efficient than *MADn*.

1.4 Objectives of Research

The goal of this study is to

1) Detect treatment effects by simultaneously controlling Type I error and the power of the test in one-way independent group design under skewed distributions using modified S_1 and *MOM-H* statistics
2) To compare the new methods with the frequently used classical methods such as ANOVA for the parametric approach and Kruskal-Wallis for the non-parametric approach.

1.5 Significance of Study

This research will contribute towards the knowledge development in the experimental design methodology especially in the experimental sciences field. Statisticians are aware that experimental design methodology depends on the assumptions of normality and treatment groups having equal variances. However, in the real world, the data are not always distributed normally. The benefit of this research is that with these new methods (produced at the end of this project), researchers (in various field, especially the experimental sciences) will not be constrained with all the assumptions mentioned earlier. They (researches) can work with the original data without having to worry about the shape of the distributions. The latest applications which are similar to this type of research can be found in the health related field(Walters and Campbell, 2004; Draghici et al., 2003). A list of abstracts on detecting treatment

effects on health issue can be found in the website:

http://www.isoqol.org/Conference/2001/Abstracts_Listing_for_Journal2001.htmlhttp://www.isoqol.org/Conference/2001/Abstracts_Listing_for_Journal2001.html

2.0 LITERATURE REVIEW

In testing central tendency (location) measures for more than two groups, the classical method, ANOVA, is among the most commonly used statistical methods in the one-way independent group design. However, this method is adversely affected by non-normality, particularly when variances are heterogenous and group sizes are unequal (Lix and Keselman, 1998). Violating the assumptions associated with this method will cause the Type I error to be disrupted. The Type I error rates will be inflated from the nominal value and power rates can be substantially reduced from the theoretical value. These liberal values of Type I error rates will subsequently result in spurious rejections of the null hypotheses. Even though it is well established that the conventional ANOVA for comparing means is not robust if the homogeneity assumption does not hold (Wilcox et al., 1986), the *F*-test in ANOVA, for example, is often employed in statistical practice even when the data suggest that population variances are unequal (Kulinskaya et al., 2003). However, consequences of the effects of these violations for test statistics are hard to gauge, and are thus important issues that need further investigation.

In the effort to overcome the sensitivity of these procedures to the violations of the assumptions, researchers in this area have sought to find alternative methods. Cochran (1937) suggested weighting the terms in the sum of squares explained by the respective inverses of the sample variances, and he provided a chi-square test for equal means based on a transformation of the *F*-test for ANOVA. However in this case, the design has to be balanced. For unbalanced design, James (1951) and Welch (1951) had suggested weighting the terms in the sum of squares explained by

estimates of the inverses of the variances of the respective sample means. This weighted sum of squares has an approximate chi-square distribution under the null hypotheses of equal population means for large sample sizes. Even if the problem of unequal variances could be overcome, the assumption of normality will always be associated with ANOVA. Even though ANOVA is known to be robust to small deviations from normality, to what extent can this method hold is unknown as there is no exact measurement of the violation or deviation from normality that we can base on, unless the sample size is big enough to guarantee the normality of sample means. This problem is common in the experimental sciences where measurements are typically repeated only a few times thus yielding small sample sizes, which of course will violate the normality assumption. Any violations, be it the non-normality or heterocedasticity, will always have some impact on the result of the ANOVA as well as the *t*-test.

In order to achieve a good test of a statistical hypothesis, Type I error rates need to be control at the nominal level. As alternatives to the ANOVA , one can seek methods from the less powerful nonparametric statistics but these less powerful methods need large sample sizes to increase power. Centering on parametric models, but not entirely convinced that the assumption is true, robust statistical methods will give ways of finding practical solutions in statistics. With high speed computers, it is now possible to apply robust statistical methods that were heretofore impractical to use.

Robust statistical methods offer useful and viable alternatives to traditional analytic methods, often yielding greater statistical power and increased sensitivity. These methods were also proven to be able to control the Type I error rates at the

nominal level (Keselman et al., 2002; 2004b; Othman et al., 2004; Syed Yahaya et al., 2004a; 2004b; Wilcox et al., 2001; Wilcox et al., 1988).

Luh and Gou (1999) agreed that approximate tests are well known alternatives for dealing with the problem of heteroscedasticity. Nevertheless they noted that although these tests are known to be the most valid tests under various conditions of heteroscedasticity investigated (Wilcox, 1989), they could not simultaneously handle the problem of non-normality. They cited the case of the Welch test (Welch, 1951) and the James second-order test (James, 1951) which were though the most valid tests under various conditions of heterogeneity investigated were nevertheless affected by non-normality conditions (Keselman et al., 1995).

Babu et al. (1999) proposed the S_1 method that can handle the problems of non-normality and heteroscedasticity simultaneously in an adaptive test setting. This method needs no trimming or transforming and will be selected when the data are skewed.

Another way of dealing with skewed data is by trimming the tail of the distribution. Working with actual data, Wilcox et al. (2000) found that power can be greatly increased by comparing trimmed means versus means and control over the probability of a Type I error can be better. However, there are practical concerns regarding trimmed means. The utmost concern is that by assumption, the amount of trimming is fixed prior to analyzing the data. The next concern is that trimming is typically assumed to be symmetric. Given these concerns, the question is how can we determine the best percentage of trimmings especially when the distribution is skewed?

In dealing with the problem of predetermined amount of trimming, Wilcox et al. (2002) suggested modified one-step *M*-estimator (*MOM*) which addresses the

problems with trimmed means. For example, if sampling is from a light-tailed distribution or normal distribution, it might be desirable to trim very few observations or perform no trimming at all. If the distribution is skewed, a natural reaction is to trim more observations from the skewed side of the empirical distribution. This central tendency estimator, like trimmed mean, can be applied to test statistics to investigate the equality of central tendency measures across treatment groups (Keselman et al., 2002; Othman et al., 2004). By using a statistic mentioned by Schrader and Hettsmansperger (1980), examined by He et al. (1990) and discussed by Wilcox (1997), Othman et al. (2004) proposed a method known as *MOM-H* which uses *MOM* as the central tendency measure.

In searching for the alternative approach in testing central tendency measures in the one-way independent group design, we suggested two robust procedures which were proposed by Babu et al. (1999) and Othman et al. (2004), i.e. the S_1 by itself, not in the adaptive and *MOM-H* respectively and combined these statistics with some selected robust scale estimators. Based on the proposed robust scale estimators by Rousseeuw and Croux (1991, 1993), three scale estimators which have the highest breakdown point and bounded influence function were selected. Two of the scale estimators, S_n and Q_n, were proposed in 1991, and the third scale estimator, T_n was proposed in 1993. Apart from these scale estimators, we were also interested in MAD_n, based on its good robust properties and for being one of the most popular robust scale estimators. The integration of these scale estimators with S_1 generated good control of Type I error when tested on a moderately skewed distribution (Syed Yahaya et al., 2004a; 2004b).

Before going in depth into the discussion on the two statistics and the selected scale estimators, the following sections will be defining some of the terminologies and the literature survey that were being used frequently through out this thesis.

2.1 Type I Error

The aim of this study is to look at the effect of Type I error and power when the problems of non-normality and heteroscedasticity occur. Type I error, α, is defined as the probability of rejecting a true null hypothesis. Since it refers to the rate of rejecting a "true" null hypothesis, therefore, it should be of a relatively small value. The null hypothesis for testing the equality of central tendency measures is given as

$$H_0 : \theta_1 = \theta_2 = ... = \theta_j ,$$

where θ_i is the central tendency parameter for F_i: i = 1, 2, …, j, and F_i is the distribution for group i.

Type I error rate is easy to access, because it involves calculating the proportion or percentage of significant statistical test (e.g. t's and F's) when the underlying population means are the same. When assumptions are met, the proportion of significance should come close to the set significance level.

In the context of hypothesis testing, the aspect of robustness is the ability of a procedure to control the Type I error rate of a test close to the nominal value (significance level), i.e. α, and stable over a range of distributions even with some deviations from its assumptions and Tiku et al. (1986) referred to this as "robustness of validity". Robust statisticians are looking for test procedures which are able to control the Type I error rates at the nominal value. By convention, a procedure can be considered robust if its Type I error is in between $0.5\alpha \le \hat{\alpha} \le 1.5\alpha$. For the nominal level $\alpha = 0.05$ the Type I error rate should be in between 0.025 and 0.075.

Empirical Type I errors rates above 0.075 are considered liberal and those below 0.025 are considered conservative. However, Guo and Luh (2000) considered a test to be robust if its empirical Type I error rate does not exceed 0.075 for the 5% level of significance used.

Type I error control is affected by the extreme conditions of non-normality and variance heterogeneity. In their investigation on the robustness of Student's *t*-test, Sawilowsky and Blair (1992) found that distributions with the extreme degree of skewness (e.g. $\gamma_1 = 1.64$) affected the Type I error control of the independent sample *t* statistic. Apart from these two problems, rates of Type I error can also be subjected to the unbalanced design, and even the pairings of unbalanced sample sizes with the unbalanced group variances. It is well known that the combination of larger variance with smaller sample size will disrupt the Type I error (Spector, 1993).

2.2 Power of a Statistical Test

The power of a statistical test is the probability of correctly rejecting a false null hypothesis, that is the probability that the test will conclude that the phenomenon exists (Cohen, 1988). Power is defined as $1-\beta$, where β is the Type II error probability. Type II error is the probability of failing to reject the null hypothesis when it needs to be rejected in favor of the alternate hypothesis. If the power of an experiment is low, then there is a good chance that the experiment will be inconclusive. Hence it is important to consider power in the design of experiments. However, in most of the work on robustness of tests, the level of the test (robustness of validity) was accorded prominence while power analysis, which is also known as "robustness of efficiency" in the robustness aspect, continued to be ignored until recently. As recent as 1997, a methodological study has found that the power of

statistical tests were not taken into account by researchers and that they continued to run a high risk of Type II error (Clark-Carter, 1997). Cohen (1988) has suggested that the neglect of power analysis exemplifies the slow movement of methodological advance. Neglect of power not only decreases the recognition of interesting effects (Type II error), but it also has a negative effect on the ability of researchers to establish statistical consensus through replication. Ottenbacher (1996, p274) points out that,

> "…*The apparently paradoxical conclusion is that the more often we are well guided by theory and prior observation, but conduct a low power study, the more we decrease the probability of replication… The responsible investigator must be concerned with statistical power. A concern with power, however, cannot end with its calculation. Because the ability to detect treatments must be optimized, the responsible scientist must also be concerned with factors that determine effect size*…"

Most studies on power deal with the calculation of power for parametric statistics where normal theory assumptions are required, for example, the *t*-test and *F*-tests. The calculations of power for robust statistics or nonstandard nonparametric statistics are not addressed at a practical level. For example, the most sought after tome on power by Cohen (1988) concentrates mainly on ANOVA and regression models and some standard nonparametric tests such as the chi-square test. What is not addressed is how violations of normality assumptions affect power estimates. Our study on power focused on the violations of normality assumptions by assuming that the factors affecting power are kept constant except for the effect size. The

power of a statistical test depends upon three parameters: i) the significance criterion, ii) the sample size, and iii) the effect size (Cohen, 1988).

2.2.1 The Significance Criterion

The significance criterion represents the standard of proof that the phenomenon exists, or the risk of mistakenly rejecting the null hypothesis. Denoted by α, it is known as the Type I error rate. The more conservative the significance level, the lower the power. Thus, using the .01 level will result in lower power than using the .05 level.

The directionality of the significance criterion also gives some impact to the power of a statistical test (Cohen, 1988). When no direction is specified, the resulting test will have less power than the test with the same α value which is directional, as long as the effect is in the expected direction.

2.2.2 The Sample Size

The reliability of sample result is always dependent upon the size of the sample. All things being equal, the larger the sample size, the greater the reliability or precision of the results, thus the greater the probability of detecting a non-null state of affairs, that is, the phenomenon under test can manifest itself more clearly against the background of variability. By increasing the sample size, the statistical power will increase.

2.2.3 The Effect Size

Cohen (1988) defined "effect size" as the degree to which the phenomenon is present in the population, or the degree to which the null hypothesis is false in relation to the alternate hypothesis**.** The null hypothesis always means that the effect size is zero**.** Specifically, in using the *t*-test for two independent groups, effect size is simply the difference between the two averages divided by the standard deviation i.e. the standardized mean difference. In the *F*-test for two or more population means, the "effect size" is the standard deviation of standardized means. Effect size measures provide a standardized index of how much impact treatments actually have on the dependent variable. Conventionally, the measures of effect size can be categorized into small, medium, and large effects depending on the on the area of research. The values are arbitrary, but the conventional definitions of effect size by Cohen (1988) are given in Table 2.1:

Table 2.1: Conventional effect size values by Cohen (1988)

# of Groups / Effect size	≥ 2 Groups
Small	0.10
Medium	0.25
Large	0.40

Several other factors such as variance and population distribution can affect power. Increasing the variance will lower the power of a statistical test. A homogenous population reduces the variance thus increasing power. With regard to population distribution, deviations from the assumption of normality usually lower the power. The type of statistical procedure used can also have some impact on power. Some of the distribution free tests are less powerful than other tests when the distribution is normal but more powerful when the distribution is highly skewed.

The main purpose of power analysis is to provide a guide in the choice of sample size. The experimenter need to specify the power he or she would like to achieve before the sample size can be estimated, but the calculation of power depend on the size of the effect in the population, and estimating the effect size is the most difficult step in power calculation. The effects can be obtained from published studies as a guide, but there is a need for caution, however, since there is a tendency for published studies to contain overestimates of effect sizes. Previous work might not be sufficiently similar to a new study to provide a valid basis for estimating the effect size**.** For this instance, it is possible to specify the minimum effect size that is considered important.

To reiterate, our study on power analysis focused on the non-normality assumption and other aspects covered in Type I error, with extra attention on effect size. The entire test done on Type I error will be repeated on each effect size. The bootstrap technique can be useful for exploring how statistical power is affected by non-normality. Beran (1986) provided mathematical and simulation results that show that a statistical test for a null hypothesis can be constructed by bootstrapping the null distribution for the test statistic. Beran (1986) also proved that the power of the test against an alternative could itself be estimated by simulation. Additionally, the uniform consistency of these simulated power functions confirms Beran's mathematical proof. Wilcox and Keselman (2002), Wilcox et al. (1998), Othman et al. (2004), Liu and Singh (1997) are among those who use bootstrapping to obtain power (for further elaboration, refer to Section 3.3). Luh and Guo (1999) suggested that another way to deal with low power due to non-normality is to replace the mean with a resistant measure of location.

Taking the above suggestions into consideration, this study combines all the aforementioned ideas in the search for a good robust method. First, we identified some of the latest robust statistics, and found two statistics, S_1 and *MOM-H*, which uses median and modified one-step *M*-estimator (*MOM*) respectively as their central tendency measure. Median and *MOM* are among the most robust central tendency measures with the highest breakdown point possible.

3.0 METHODOLOGY

In this study, S_1 and *MOM-H* will be modified using the alternative robust scale estimators S_n, T_n, and MAD_n. Listed below are all the procedures that will be employed:

1. S_1 with $\hat{\omega}$
3. S_1 with T_n
4. S_1 with MAD_n
5. *MOM-H* with S_n
6. *MOM-H* with T_n
7. *MOM-H* with MAD_n.

S_1 with $\hat{\omega}$ is the original procedure for S_1. For the purpose of comparison, it is included in this study. These procedures will be compared to the most frequently used classical methods such as ANOVA for the parametric and Kruskal-Wallis for the non-parametric approach.

3.1 Variables Manipulated

Since this study deals with robust methods where sensitivity to non-normality and variances heterogeneity are of the main concern, manipulating variables could help in identifying the sturdiness or robustness of each procedure. In the following subsections, we will discuss the variables that are manipulated to create conditions which are known to highlight the strengths and weaknesses of tests designed to determine the central tendency measures equality namely,

- Types of distributions
- Variance heterogeneity

- Nature of pairings

However, this study only focused on the 4 groups unbalanced design with $n_1 = 10, n_2 = 15, n_3 = 25, n_4 = 30$.

3.1.1 Types of Distributions

The two major problems in ANOVA are the violations of the normality and variance homogeneity. Under non-normal distributions, these statistics performed quite poorly (Bradley, 1968) in terms of Type I error. A slight departure from normality had a great negative effect on power (Sawilowsky and Blair, 1992; Wilcox, 1995).

In investigating the effects of distributional shape on Type I error and power, three types of distributions representing different levels of skewness were considered. The standard normal distribution represents distribution with zero skewness. Apart from this, two non-normal distributions were also analyzed. They were the chi-square distribution with three degrees of freedom (chi-square) and the *g*-and-*h* distribution with g = 0.5 and h = 0.5 (*g*-and-*h*). The chi-square distribution (χ_3^2) was chosen to represent mild skewness and the *g*-and-*h* distribution (Hoaglin, 1985) to represent extreme skewness. The skewness and kurtosis values for the χ_3^2 distribution are γ_1 = 1.63 and γ_2 = 4.00 respectively (Othman et al., 2004). On the other hand, the respective theoretical values for skewness and kurtosis of the *g*-and-*h* distribution were g = 0.5 and h = 0.5 are $\gamma_1 = \gamma_2 =$ undefined**.** The purpose of selecting these extreme values is based on the assumption that if a method performed well under seemingly large departures from normality, then it can be safely assumed that the

same method will also perform well for distributions of lesser skewness (which are normally encountered in practice).

3.1.2 Variance Heterogeneity

Variance heterogeneity is one of the two major problems researchers always encounter when testing the equality of location measures. The classical *F*-test for unequal means in a one-way ANOVA has been known to yield misleading results when there exist different population variances (Kulinskaya et al., 2003). For instance, if sample sizes are equal and the variances are slightly heterogeneous, the one-way ANOVA inflates to a lesser degree (Box, 1954; Sawilowsky, 1990). With moderate (1:1:6) or large (1:1:12) heterogeneous variances, however, the one way-ANOVA inflates to a greater degree, even with sample sizes being equal (Rogan and Keselman, 1977; Tormarkin and Serlin, 1986)

To investigate the effect of variance heterogeneity on Type I error rates and power, variances with a 36:1 ratio were assigned to the groups. Though this ratio may seem large, ratios larger than that assigned in this study have been reported in the literature (Keselman et al., 2004a). After reviewing articles published in prominent education and psychology journals, Keselman et al. (1998) noted that they found ratios as large as 24:1 and 29:1 respectively in one-way and in completely factorial randomized designs. Wilcox (2003) cited data sets where the ratio was 17,977:1 as mentioned in Keselman et al. (2004a), Thus, although the ratio of 1:36 may seem large, it still appears to be a reasonable figure with which investigations into how well the tests perform under potentially extreme conditions can be confidently undertaken. The reason being that if a procedure works under an extreme degree of

heterogeneity, it is likely to work under most conditions of heterogeneity which are likely to be encountered by researchers.

3.1.3 Nature of Pairings

When unequal variances were paired with unequal sample sizes, negative and positive pairings were formed. A positive pairing involves the pairing of the largest number of group observations with the largest group variance, and the smallest group observations with the smallest group variance. For the negative pairing, the largest group observation was paired with the smallest group variance, while the smallest group observation was paired with the largest group variance (refer to Table 3.1)

Table 3.1: Design specification for the unbalanced $J = 4$

Pairing	**Group Sizes**				**Group Variances**			
	1	2	3	4	1	2	3	4
Positive	10	15	25	30	1	1	1	36
Negative	10	15	25	30	36	1	1	1

The nature of pairings of sample sizes and variances do have an effect on Type I error (Keselman et al., 1998; Keselman et al., 2004b; Othman et al., 2004). For instance, in a pairing wherein there is an unequal number of group observations and moderate or large heterogenous variances, the inflations of Type I error become more pronounced in comparison to those from equal sample sizes. These inflations are further exacerbated if the group observations with the largest *n* has the smallest variance, and the group observations with the smallest *n* has the largest variance (Box, 1953; Snedecor and Cochran, 1980; Spector, 1993). Empirically, the positive and negative pairings typically produce conservative and liberal results, respectively (Othman et

al., 2004). Therefore, to evaluate the robustness of the procedures in relation to the nature of the pairings, each of the proposed procedures was examined under the two types of pairings.

3.2 Data Generation

This study is based on simulated data in which the programs and simulations were run using the SAS/IML Version 8. The study on distributional shape required the simulation of data according to the types of distribution. In terms of the data generation procedure, the following will explain how the pseudo-random variates for each particular distributional shape were obtained:

a) Standard normal distribution

   This involved the straight forward usage of SAS generator RANNOR (SAS Institute, 1999) with mean equaling 0 and standard deviation equaling 1.

b) Chi-square distribution with three degrees of freedom.

   i. Generate three standard normal variates using (a)

   ii. Square each of the three standard normal variates

   iii. Total them up.

c) *g*-and-*h* distribution with $g = h = 0.5$

   i. Generate a standard normal variates, Z_{ij}, using (a)

   ii. Convert the standard normal variates to random variables via equation

$$Y_{ij} = \begin{cases} \dfrac{\exp(gZ_{ij}) - 1}{g} \exp(hZ_{ij}^{\,2}/2), & g \neq 0 \\ Z_{ij} \exp(hZ_{ij}^{\,2}/2), & g = 0 \end{cases} \qquad [3.1]$$

The parameter g controls the amount of skewness, while parameter h controls the kurtosis. For the case of $g = 0$ and $h = 0$, Y_{ij} = Z, corresponds to a standard normal distribution. It was noted that the tails of the distribution became heavier as h increased and were further skewed as g increased. For this study, we chose $g = h = 0.5$ to create a condition of extreme non-normality.

In general, when dealing with skewed distributions, the central tendency measures such as the median and *MOM* have values unequal to zero. To ensure that the null hypothesis remains true, the observations, Y_{ij}, from each simulated skewed distributions were shifted (standardized) by subtracting the population central tendency parameter (θ) from the observations such that,

$$X_{ij} = Y_{ij} - \theta \qquad [3.2]$$

The values of θ are determined by computing $\hat{\theta}$ with one million observations generated from the distribution under study (Othman et al., 2004; Wilcox and Keselman, 2003). Specifically, when dealing with median, the population median should be subtracted from Y_{ij} in order to ensure that the null hypothesis for equal population medians remains true. Similarly, when dealing with *MOM*, the population *MOM* should be subtracted for the same aforementioned reason. Based on the million observations generated, the population median and population *MOM* corresponding to the types of distributions were recorded as in Table 3.2,

Table 3.2: Location parameters with respect to distributions

Distributions	**Location Parameters**	
	Median	***MOM***
Normal	0	0
Chi-square (3df)	2.366	2.50
g-and-*h* (*g*=0.5, *h*=0.5)	0	-0.025

The standardized (shifted) variates were then transformed to suit the study conditions. For the case of unequal variances, to obtain a distribution with a standard deviation σ_j, each X_{ij} from Equation [3.2] was multiplied by the square root of σ_j^2 as follows,

$$X_{ij} = (Y_{ij} - \theta) \times \sqrt{\sigma_j^2} \qquad [3.3]$$

For example, if $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$ = (36, 1, 1, 1), we needed firstly to standardize every observations in group 1 according to the central tendency parameter being investigated before multiplying the standardized observation with 6.

In the analysis of the Type I error rates, the groups central tendency measures were set to be $(\theta_1, \theta_2, \theta_3, \theta_4)$ = (0, 0, 0, 0) for J = 4 case. However, for the power analysis, these values vary according to the suggested effect size and pattern variability. For example, when J = 4 under the unbalanced design, the setting of the central tendency measures for the intermediate pattern variability was (-1, -1, 1, 1). To conform with either Type I error or power, θ_j was added to Equation [3.26] such that,

$$X_{ij} = (Y_{ij} - \theta) \times \sqrt{\sigma_j^2} + \theta_j \qquad [3.4]$$

When testing for Type I error rates, the value for θ_j is always set at zero while for power the value depends on the settings of the central tendency measures. For example, if we want to investigate on the Type I error rate of the test for the equality of medians, given that the distribution is mildly skewed; first, the mildly skewed distribution, which was represented by the chi-square (with 3 degrees of freedom) distribution, is generated. Then, the distribution is standardized by subtracting each Y_{ij} with the population median (θ = 2.366), so that its median became zero. For the distribution to have a variance of σ^2 = 36 for example, the standardized observations were later multiplied by the square root of 36. Then the assigned central tendency measure, e.g. θ_1 = 0 is added to each of the following observations:

$$X_{ij} = (Y_{ij} - \theta) \times \sqrt{\sigma_j^2} + \theta_j$$

If $\theta = 2.366$, $\sigma_1^2 = 36$, and $\theta_j = 0$, then

$$X_{ij} = (Y_{ij} - 2.366) \times \sqrt{36} + 0$$

As for the power analysis, the setting of groups' central tendency measures depends on the patterns of variability as suggested by Cohen (1988). In this case, at least one of the assigned central tendency measures will not be equal to zero. The following section will discuss on the settings of central tendency measures for power analysis. The general procedure in this section was to generate observations from the three distributions, which represent different levels of skewness combined with unbalanced sample sizes and unequal variances or different levels of skewness combined with balanced sample sizes and variances all equaling one.

For each condition, 5000 datasets were simulated using a 0.05 statistical significance level ($\alpha = 0.05$). According to Manly (1997), for a test at 5% level of significance, the minimum 1000 datasets are almost certain to yield the same results

as would a full distribution. However, when using 5000 datasets, better sampling limits within which estimated significance levels will fall 99% of the time were obtained when compared to the use of 1000 datasets (Manly, 1997). Therefore, based on this finding, this study chose 5000 datasets as the number of randomizations. Each of these simulated datasets was then bootstrapped 599 times.

3.3 Bootstrap Method

The bootstrap was introduced by Efron (1979) as a computer-based method for estimating the standard error of $\hat{\theta}$. This method has gained a great deal of popularity in empirical research. The word bootstrap is used to indicate that the observed data are used not only to obtain an estimate of the parameter but also to generate new samples from which many more estimates may be obtained, and hence an idea of the variability of the estimate (Staude and Sheather, 1990). This method treats the sample like the population and draws samples from this pseudo population in order to assess

(a) variability of an estimator

(b) bias of an estimator

(c) predictive performance of a rule

(d) significance of a test.

The beauty of the bootstrap is in its

i. wide applicability – requires no theoretical calculation, and is available no matter how mathematically complicated the estimator may be.

ii. increased accuracy

iii. ability to take advantage of modern computing and completely automatic.

The basic idea is that in the absence of any other information about a population, the values in a random sample are the best guide to the distribution, and resampling the sample is the best guide to what can be expected from resampling the population. To measure the reliability of the inferences made, statistician must understand the distribution of any statistics that are used in the analysis. In situations where a well understood statistic is used, such as the sample mean, this is easily done analytically. However, in many applications, we do not want to be limited to using such simple statistics or to making simplified assumptions.

To conduct a simulation experiment, a model that represents the population or phenomena of interest and a way to generate random numbers (according to the model) using a computer is needed. The data generated from the model can then be studied as if they were observations.

4.0 RESULTS

According to the Bradley's liberal criterion of robustness (Bradley, 1978), a test can be considered robust if its empirical rate of Type I error, $\hat{\alpha}$, is within the interval $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. In this study, the nominal level is set at $\alpha = 0.05$, thus, the empirical Type I error rate should be in between $0.025 \leq \hat{\alpha} \leq 0.075$. Our aim is to identify those procedures that are able to control the rate of Type I error within the bounds of robustness, i.e. within the 0.025 to 0.075 range.

Even though the main focus of this study is to compare the Type I error rates of the procedures investigated, findings on the power rates were also included. For a test to be considered robust, it must possess the ability to control its Type I error rate. In addition to a good control of Type I error rates, a test will be more convincing if it can generate high power rates.

4.1 Type I Error Rates

Based on the Bradley's criterion of robustness, we compare the Type I error rates from the seven robust procedures with ANOVA and Kruskall Wallis as shown in Table 4.1

4.1.1 Type I Error Rates Under Normal Distribution

The second column in Table 2 displays the Type I error rates for all the procedures investigated when the data is normally distributed. Our purpose in this study is to compare the Type I error values with the nominal $\alpha = 0.05$..

Table 4.1: Empirical Type I error rates

Procedures	**Distributions**					
	N(0,1) Normal		χ^2_3 **Mildly Skewed**		**g = 0.5 , h = 0.5 Extremely Skewed**	
	+ve	**-ve**	**+ve**	**-ve**	**+ve**	**-ve**
***MOMH_T*$_n$**	0.0486	0.0542	0.0694	0.0650	0.0286	0.0316
Average	**0.0514**		**0.0672**		**0.0301**	
***MOMH_S*$_n$**	0.0478	0.0540	0.0642	0.0642	0.0268	0.0308
Average	**0.0509**		**0.0642**		**0.0288**	
***MOMH_MAD*$_n$**	0.0486	0.0520	0.0646	0.0660	0.0292	0.0286
Average	**0.0503**		**0.0653**		**0.0289**	
S$_1$_T$_n$	0.0244	0.0260	0.0264	0.0330	0.0174	0.0194
Average	**0.0252**		**0.0297**		**0.0184**	
S$_1$_ S$_n$	0.0214	0.0254	0.0234	0.0284	0.0152	0.0192
Average	**0.0234**		**0.0259**		**0.0172**	
S$_1$_ MAD$_n$	0.0248	0.0268	0.0236	0.0324	0.0188	0.0206
Average	**0.0258**		**0.028**		**0.0197**	
***S$_1$_* $\hat{\omega}$**	0.0278	0.0302	0.0246	0.0278	0.0078	0.0102
Average	**0.029**		**0.0262**		**0.009**	
Kruskall Wallis	0.0448	0.1158	0.0492	0.1180	0.0498	0.1022
Average	**0.0803**		**0.0836**		**0.076**	
ANOVA	0.0336	0.2850	0.0526	0.2976	0.1492	0.3554

Average	**0.1593**	**0.1751**	**0.2523**

The closer the values to $\alpha = 0.05$, the better the performance of the procedures. By referring to the average Type I error rates (in bold), the values closer to 0.05 are all produced by the *MOM-H* procedures namely *MOMH_*T_n , *MOMH_*S_n, and *MOMH_MAD*$_n$. Even the poorly performed S_1 procedures are still able to control the Type I error rates within the Bradley's robust criterion. Nonetheless, ANOVA, which is the most frequently used method when the data is normally distributed, fares the worst. Under this condition, the only violated assumption for ANOVA is unequal variances, whereas the distribution assumption is fully obeyed. Under usual practice, when the ANOVA assumptions are violated, the next best alternative is to use a non parametric procedure such as Kruskall Wallis. However, the values produce by this method are also out of control albeit all the assumptions are fully obeyed.

The large values signify that the methods are unable to handle extreme data, as reflected in the average values for the two conventional methods. These conventional methods do not encounter any problem in controlling Type I error rates when the pairing is positive, but not for the negative pairing. The Type I error rates inflate tremendously when the sample sizes and variances is negatively paired.

4.1.2 Type I Error Rates Under Mildly Skewed Distribution

The performance of the procedures under mildly skewed distribution is shown in the third column of Table 2. All the Type I error rates for the proposed procedures, i.e. *MOM-H* and S_1 are within the Bradley's robust limit. Even though the values for the *MOM-H* procedures enlarge and digress from the nominal value, the procedures are still among the best. The procedures for S_1 improves when the values increase,

and regress towards α = 0.05. However, the performance for the conventional methods, ANOVA and Kruskall Wallis deteriorate with very liberal values especially for ANOVA. Furthermore, the disparity in Type I error values between the positive and negative pairings are more noticeable in the two conventional methods. The values are not consistent with respect to the sign of the pairings.

4.1.3 Type I Error Rates Under Extremely Skewed Distribution

The last column of Table 4.1 presents the Type I error rates for all the procedures under extreme skewness. Only *MOM-H* procedures can be considered robust according to the Bradley's robust criterion. All the Type I error rates for the S_1 procedures deflate far below the minimum limit of Bradley's. The original S_1 procedure produced the most conservative Type I error value i.e. 0.009. Even though the conventional Kruskall Wallis exhibit some improvement in the performance, the Type error rate of 0.076 is beyond the Bradley's maximum limit of robustness. In the case of ANOVA, the resulted Type I error rate of 0.2523 is too far beyond the acceptance region, thus making it the worst of all methods tested. Even the positive pairing for ANOVA is too liberal to be considered robust under this condition unlike the other two conditions; normal and mildly skewed.

4.2 Power Rates

Table 4.2 provides the results of the power rates for all the procedures investigated.

By convention, a test can be considered powerful when the power rate reaches the 0.8 level. As shown in Table 3, all the *MOM-H* procedures generate low power rates

In contrast, the S_1 procedures show improvement when the power rates approaching the 0.8 value in the case of normal and extremely skewed distributions. However, above all are the values for the Kruskall Wallis test. With the exception of the mildly skewed distribution, the power rates for the other distributions using Kruskall Wallis test are above the 0.8 level. This shows that the conventional non parametric method is a powerful test.

Table 4.2: Empirical Power Rates

Procedures	**Distributions**					
	N(0,1) Normal		χ_3^2 **Mildly Skewed**		**g = 0.5 , h = 0.5 Extremely Skewed**	
	+ve	**-ve**	**+ve**	**-ve**	**+ve**	**-ve**
$MOMH\_T_n$	0.2966	0.2212	0.0720	0.1126	0.1520	0.1396
Average	***0.2589***		***0.0923***		***0.1458***	
$MOMH\_S_n$	0.3032	0.2120	0.0744	0.1148	0.1472	0.1392
Average	***0.2576***		***0.0946***		***0.1432***	
$MOMH\_MAD_n$	0.2878	0.2046	0.0726	0.1174	0.1496	0.1340
Average	***0.2462***		***0.0950***		***0.1418***	
$S_1\_T_n$	0.7618	0.5058	0.1776	0.1666	0.7342	0.5340
Average	***0.6338***		***0.1721***		***0.6341***	
$S_1\_ S_n$	0.7476	0.4426	0.1568	0.1362	0.7028	0.4594
Average	***0.5951***		***0.1465***		***0.5811***	
$S_1\_ MAD_n$	0.7506	0.4260	0.1594	0.1284	0.7236	0.4698
Average	***0.5883***		***0.1439***		***0.5967***	
$S_1\_ \hat{\omega}$	0.8936	0.8124	0.2100	0.2000	0.5954	0.4330
Average	***0.8530***		***0.2050***		***0.5142***	
Kruskall Wallis	0.9980	0.9994	0.4458	0.7250	0.8020	0.9054

Average	***0.9987***		***0.5854***		***0.8537***	
ANOVA	0.2714	0.8696	0.0510	0.4066	0.1050	0.4892
Average	***0.5705***		***0.2288***		***0.2971***	

On the contrary, the result for the most frequently used conventional parametric test, ANOVA, is disappointing. Even under normal distribution, the test is not powerful. This could be due to the violation of the homogeneity of variances assumption.

In the case of the pairing of group sizes and variances, the power rates generated from the robust procedures show higher rates when the pairing is positive, but the result contradicts with the other two conventional procedures. For Kruskall Wallis and ANOVA, the rates are higher when the group size and variances were negatively paired.

5.0 CONCLUSION

The violation of assumptions such as non normality and heteroscedasticity (unequal variances) caused ANOVA to loose control of its Type I error. Thus, when non normality is suspected, the next alternative to ANOVA is the non parametric method, such as Kruskall Wallis. Kruskall Wallis test makes no assumptions about the distribution of the data and the equality of variance. Since this test does not make a distributional assumption, it is not as powerful as the ANOVA. In recent years, investigators are searching for an estimator that is powerful, stable and insensitive to all the assumptions. Robust statistics combine the virtues of both, the parametric and the non-parametric approach**.** The approach in robust inference is different from the non parametric (eg. Kruskal Wallis) wherein there is a working assumption about the form of the distribution, but we are not entirely convinced that the assumption is true.

This study shows that the average Type I error rates across distributions for ANOVA and Kruskall Wallis are liberal and none are robust according to the criterion for robustness (i.e. $0.025 \leq \hat{\alpha} \leq 0.075$) as compared to the proposed robust procedures. These Type I error rates worsen when the sample sizes and variances are negatively paired. Under extreme condition, ANOVA fares the worst with $p = 0.2$. In

contrast, all the procedures for the proposed *MOM-H* are robust even under extreme condition. *MOM-H* with T_n generates the best Type I error rate among the other procedures under this condition. Even though the proposed S_1 procedures perform worse than *MOM-H* procedures in terms of Type I error rates, the S_1 procedures are much better than the ANOVA and Kruskal Wallis under normal distribution. The proposed procedures are also proven to be robust to the nature of pairings of the sample sizes and group variances unlike the other two conventional methods.

In terms of power, the proposed S_1 procedures generate moderate power rates, but not the proposed *MOM H* procedures, which produce quite low power rates. However, compensating the weakness is the good Type I error rates generated from the procedures. In statistics, a reliable and robust method depends on the value of its Type I error rate , which shows that the method is in control regardless on conditions. A high power rate corresponds to the method will be an added value. Therefore in the case of *MOM-H,* this method is proven to be able to control the value of Type I error rate even under extreme condition. Kruskal Wallis method generates the best power rates, but this method unable to sustain its Type I error under certain condition. The most frequently used ANOVA is susceptible to conditions tested. Even under normal distribution, when the variance are not equal, in addition to the unequal group sizes, the Type I error and power are greatly distorted.

From this study, a few potential alternative methods in testing for the equality of the central tendency measures under skewed distributions were identified. When symmetry is suspect, we would like to suggest using these methods especially the proposed *MOM-H* procedures as the alternative to the traditional methods.

In this study, we used bootstrapping method to test the hypotheses. The reason of using bootstrapping method was due to the fact that the sampling

distributions for the statistics used were intractable. Certainly, further research is required in arriving at the sampling distributions. Too much reliance on resampling techniques to come up with a pseudo sampling distribution goes against the grain of traditional mathematical statistics whereby a sampling distribution or an asymptotic sampling distribution is always preferable.

REFERENCES

Babu, G.J., Padmanabhan, A.R and Puri, M.L. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal*. **41**: 321-339.

Beran, R (1986). Simulated Power Functions. *The Annals of Statistics*. **14(1)**: 151-173.

Box, G.E.P. (1953). Non-normality and tests of variances. *Biometrika*. **40**: 318 – 355.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*. **25**: 290 – 302.

Bradley, J.V. (1968). Distribution-free statistical tests. Englewood Cliffs, NJ: Prentice Hall.

Clark-Carter, D.(1997). The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology*. **88**: 71-83.

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*. (suppl.4): 102-118.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York

Draghici, S., Kulaeva, O., Hoff, B., Petrov, A., Shams, S., and Tainsky, M.A. (2003). Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics*. **19**(**11**):1348-59

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". *The Annals of Statistics* **7** (1): 1–26

He, X., Simpsom, D.G., and Portnoy, L.S. (1990). Breakdown Robustness of Test. *Journal of American Statistical Association*. **85**: 446-452.

Hoaglin, D.C. (1985). Summarizing shape numerically: The *g*-and-*h* distributions. In D. Hoaglin, F.Mosteller, and J. Tukey (eds.), *Exploring Data Tables, Trends, and Shapes*. Wiley, New York

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

http://www.isoqol.org/Conference/2001/Abstracts_Listing_for_Journal2001.htmlhttp://www.isoqol.org/Conference/2001/Abstracts_Listing_for_Journal2001.html

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*. **38**: 324-329.

Keselman, H.J., Carriere, K.C., and Lix, L.M. (1995). Robust and powerful nonorthogonal analyses. *Psychometrika*. **60**: 395-418.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B, Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., and Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*. **68(3)**: 350-386.

Keselman, H.J., Othman, A.R., H.J., Wilcox, R.R., Fradette, K. (2004b). The new and improved two-sample t-test. *American Psychological Society*. **15(1)**: 57-51.

Keselman, H.J., Wilcox, R.R., Algina, J., Fradette, K.,Othman, A.R. (2004a). A power comparison of robust test statistics based on adaptive estimators. *Journal of Modern Applied Statistical Methods*. **3(1)**: 27-38.

Kulinskaya, E., Staudte, R.G., Gao, H. (2003). Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Communications in Statistics - Theory and Methods*. **32**: 2353-2371.

Liu, Regina Y and Singh, Kesar. (1997). Notions of limiting p-values based on data depth and bootstrap. *Journal of American Statistical Association*. **92(437)**: 266-277.

Lix, L.M and Keselman, H.J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*. **58**: 409-42.

Luh, W. and Gou, J. (1999). A Powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*. **52**: 303-320.

Manly, B.F.J. (1997) *Randomization, bootstrap and Monte Carlo. Methods in biology*, 2nd edn. Chapman and Hall, London

Ottenbacher, K.J. (1996). The power of replications and replications of power. *The American Statistician*. **50(3)**: 271-275.

Othman,A.R., Keselman, H.J., Padmanabhan, A.R., Wilcox, R.R and Fradette, K (2004) Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology*. **00**, 1 – 21.

Rogan, J.C. and Keselman, H.J. (1977). Is the ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coeffiecient of variation. *American Educational Research Journal*. **14**, 493 –498.

Rousseeuw, P.J.and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*. **88**: 1273-1283.

Rousseeuw, P.J.and Croux, C. (1991). Alternatives to the median absolute deviation. *Technical Report 91-43*, Universitaire Instelling Antwerpen, Belgium.

SAS Institute Inc. (1999). *SAS/IML User's Guide version 8*. Cary, NC: SAS Institute Inc.

Sawilowsky, S.S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*. **60(1)**: 91 – 126.

Schrader. R.M. and Hettmansperger, T.P. (1980). Robust Analysis of Variance Based Upon a Likelihood Ratio Criterion. *Biometrika,* **67(1):** 93-101.

Snedecor, G.W. and Cochran, W.G. (1980). *Statistical Methods* (7^{th}.ed.). Iowa University Press, Ames, IA.

Spector, Paul E. (1993). *SAS Programming for Researchers and Social Scientists* Sage Publication Inc., Newbury Park.

Staudte, R.G. and Sheather, S.J. (1990). *Robust Estimation and Testing*. John Wiley & Sons Inc., New York.

Syed Yahaya, S.S., Othman, A.R. and Keselman, H.J. (2004a). *Testing the equality of location parameters for skewed distributions using S1 with high breakdown robust scale estimators*. In M.Hubert, G. Pison, A. Struyf and S. Van Aelst (Eds.), Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology, Birkhauser, Basel. 319 – 328.

Syed Yahaya, S.S., Othman, A.R., and Keselman,H.J. (2004b). An Alternative Approach for Testing Location Measures in the One way Independent Group Design. Proceedings of the International Conference on Statistics and Mathematics and Its Applications in the Development of Science and Technology, Bandung, Indonesia. Oct. 4 – 6.

Tiku, M.L., Tan, W.Y., and Balakrishnan, N. (1986). *Robust Inference*. Marcel and Dekker, New York

Tomarkin, A.J. and Serlin, R.C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*. **99(1)**: 90 – 99.

Walters, S.,Campbell, M. (2004). The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36). *Health and Quality of Life Outcomes* **70 (2)**

Welch, B.L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*. **38**: 330-336.

Wilcox, R.R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. *Journal of Educational Statistics*. **14**: 269-278.

Wilcox, R.R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology*. **48**: 99-114.

Wilcox, R.R. (1997) *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, New York.

Wilcox, R.R. (2003). *Applying Contemporatry Statistical Techniques*. Academic Press, San Diego.

Wilcox, R.R. and Keselman, H.J. (2002). Power Analyses When Comparing Trimmed Means. *Journal of Modern Applied Statistical Methods*. **1(1)**: 24-31.

Wilcox, R.R., Charlin, V.L., Thompson, K.L. (1986). New Monte Carlo Results on the Robustness of the ANOVA F, W and F* Statistics. *Communications in Statistics-Simulations*, **15(4)**: 933-943.

Wilcox, R.R., Keselman, H.J., Kowalchuk, R.K. (1998). Can test for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology*. **51:** 123-134.

Wilcox,R.R., Keselman H.J., Muska, J., Cribbie, R. (2000). Repeated Measures ANOVA : Some New Results on Comparing Trimmed Means and Means. *The British Psychological Society***.** 53**: 69-82.**

Wilcox, R.R. and Keselman, H.J. (2002). Power Analyses When Comparing Trimmed Means. *Journal of Modern Applied Statistical Methods*. **1(1)**: 24-31.

Wilcox, R. R. and Keselman H.J.(2003). Repeated Measures ANOVA Based on a Modified One-Step M-Estimator. *Journal of British Mathematical and Statistical Psychology*. **56(1)**: 15 – 26.

Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*. **61**: 165-170